Utrecht Multi-Person Motion (UMPM) benchmark

N.P. van der Aa, X. Luo, G.-J. Giezeman

R.T. Tan, R.C. Veltkamp

Technical Report UU-CS-2011-027
September 2011

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

Abstract

Analyzing human motion, including tracking and pose estimation, is a major topic in computer vision for several decades. Many methods have been and will be developed further in the future. To have a systematic and quantitative evaluation of such methods, ground truth data of the 3D human motion is scientifically required. For scenes limited to only a single person, there exist some publicly available data sets like HumanEva that provide synchronized video sequences together with detailed ground truth 3D data. However, for multiple persons, such a data set currently does not exist. In this report, we present the Utrecht Multi-Person Motion (UMPM) benchmark, which includes synchronized motion capture data and video sequences from multiple viewpoints for multi-person motion including multi-person interaction. The data set is available to the research community via http://www.projects.science.uu.nl/umpm/including documentation and software to promote further research in multi-person articulated human motion analysis. This report gives a detailed literature survey on publicly available data sets, describes the details of the data acquisition, the format of the provided data and the accompanying software.

Acknowledgement

The UMPM benchmark is created by members of the Multimedia and Geometry group of Utrecht University in the Netherlands. To advance the state-of-the-art in human motion estimation, this benchmark is made available to the research community. Without prior approval from Utrecht University, the UMPM benchmark, in whole or in part, may not be modified or used for commercial purposes.

All documents and papers that report on research that uses the UMPM benchmark are requested to acknowledge the use of the data set by including an appropriate citation to the following:

> N.P. van der Aa, X. Luo, G.-J. Giezeman, R.T. Tan, and R.C. Veltkamp. Utrecht Multi-Person Motion (UMPM) benchmark: a multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *Proceedings of the Workshop on Human Interaction in Computer Vision (HICV), 2011. In conjunction with ICCV 2011.*

Thanks go out to the voluntary actors that have made the UMPM benchmark possible:

Andreea Barac
Zamarak
Sander van de Ven
Saskia Groenewegen
Casper van Wezel
Pascal van Beek
Dennis Grootendorst
Helge Giezeman
Xinghan Luo
Heinrich Krüger
Thomáš Hodaň
Geert-Jan Giezeman
Yuri Parijs
Frank Baars
Rita Janssen
Rob Schuddeboom
Nico van der Aa
Diederik Roijers
Thijs van Lankveld
Bas de Haas
Martin de Jong

Note that the subject consent permits publication (paper or web-based) of the data (including image data) for scientific purposes only.

This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO).

Universiteit Utrecht

Game research
for training and
entertainment

Contents

1 Introduction

Analyzing human motion, including tracking and pose estimation, is a major topic in computer vision for several decades. Many methods have been and will be developped further in the future. To have a systematic and quantitative evaluation of such methods, ground truth data of the 3D human motion is scientifically required. For scenes limited to only a single person, there exist some publicly available data sets like HumanEva [35, 36] that provide synchronized video sequences together with detailed ground truth 3D data. However, for multiple persons, such a data set currently does not exist. At Utrecht University, we have created the UMPM benchmark, which includes synchronized motion capture data and video sequences from multiple viewpoints for multi-person motion including multi-person interaction.

The UMPM benchmark is meant to evaluate human motion capturing for multiple subjects in a similar way as HumanEva does for a single subject. The benchmark provides four synchronized color videos as input of the articulated human motion capture method. The ground truth 3D information is available in 3 formats: (1) 37 marker positions per subject, (2) 15 joint positions obtained directly by averaging the marker positions, and (3) 15 corrected joint postions by enforcing kinematic constraints. The user may choose which one is used, but any pose recovery/tracking method has to translate its results to either one of these formats. To facilitate the use of this data set, we provide some additional material. First, to match the video sequences to the ground truth, the calibration parameters of each camera are given, namely (1) internal camera parameters, (2) external parameters, and (3) distortion parameters. Second, the color cameras are globally aligned with the motion capture cameras. Finally, for methods that rely on background subtraction to detect or track a subject, we provide the background images. The video recordings, C3D files, calibration parameters, background images, documentation and software of the UMPM benchmark can be downloaded from http://www.projects.science.uu.nl/umpm/.

This technical report describes the details of the UMPM benchmark, namely

- description of other publicly available datasets (Section 2)
- detailed description of the acquisition of the UMPM benchmark (Section 3)
- detailed description of the provided data (Section 4)
- future extensions (Section 5)

2 Related work

In recent decades a lot of research in the field of computer vision has been validated by specially designed benchmarks. To validate algorithms on accuracy and speed, benchmark datasets have been set up in different application areas such as the FERET database for face recognition [5] and the Middlebury Stereo Datasets for stereo vision [6]. The goal of these benchmarks is that researchers have a common benchmark for their field of interest and can compare their results objectively with others. The UMPM benchmark could be used as a benchmark for multi-person human motion analysis.

To perform human motion capturing, subjects first have to be detected and tracked in an image. Since detection of humans in a natural environment is primarily performed in the area of surveillance, the benchmarks are mostly restricted to pedestrian detection. Pedestrians are typically standing upstraight and therefore restrict the pose. One of the first benchmarks for pedestrian detection was the MIT CBCL Pedestrian Database [31]. An overview of benchmarks for pedestrian detection are given in Table 1. The idea behind these benchmarks is that a training set is provided consisting of normalized images of pedestrians and non-pedestrians (sky, mountains, bycicles, etc.). The test data is provided to see how well the pedestrain detection algorithm works. These benchmarks are mostly recorded in an outdoor environment or at least in a challenging environment, where the detection is actually a problem. All of the aforementioned data sets except for the Stereo Pedestrian Detection Evaluation Dataset [26] include monocular video recordings only. A recent survey on pedestrian detection, including a description of the benchmarks of Table 1, is given in [20]. In a controlled (indoor) environment, detection of a human does not have to be an issue. The only objects changing the scenery are the persons appearing in the room. Simple background subtraction methods can already detect the person.

Keeping track of a person over subsequent frames is a requisite of many computer vision applications of which pose estimation and gesture recognition are only a subset. Robustly tracking of a human is challenging especially when multiple persons are present, since self-occlusions or occlusions by other people or objects can deteriorate the tracking results. Therefore, most benchmarks for tracking purposes include multiple cameras ranging from 2 to 8. In some benchmarks like the CAVIAR benchmark [2], wide-angle lenses are used to capture more area of the scene. The major classification in human tracking benchmarks is based on whether

Table 1: Overview of pedestrian detection benchmarks.

Name	Year
MIT CBCL Pedestrian Database [31]	1997
INRIA pedestrian dataset [18]	2005
ETH Pedestrian Dataset [21]	2007
Stereo Pedestrian Detection Evaluation Dataset [26]	2007
NICTA Pedestrian Dataset [32]	2008
Caltech Pedestrian Dataset [19]	2009
Daimler Pedestrian Detection Benchmark Dataset [20]	2009
TUD-Brussels and TUD-MotionPairs data sets [40]	2009

pedestrians are observed or not. Examples of pedestrian tracking benchmarks are the CVLAB Multi-camera pedestrian data set [23], the BEHAVE dataset [1], the SPEVI dataset [13] and the benchmarks of the PETS workshops [2, 8, 9, 10, 12]. The early benchmarks like the PETS2002 benchmark [8] were meant to track an indivual person only. The more recent ones like the PETS2009 benchmarks [10] allows you to track a person in a crowd or even track a group of people. For the non-pedestrian tracking benchmarks (such as sports), the poses are less restrictive and therefore much more challenging. To focus on pose estimation, the detection must be obvious, implying a restriction to one person only and a simple scenery, and the ground truth of the tracking should be more refined. For example, instead of finding the central point of the body, the locations of all bodyparts should be detected.

For pose estimation and gesture recognition several benchmarks have been suggested, like HumanEva [35, 36], CMU Motion of Body Database (CMU-MoBo) [24], CMU Multi-Modal Activity Database [4], TUM Kitchen data set [37], Multimodal Motion Capture Indoor Dataset (MPI08) [33], MuHAVi [7], the Keck dataset [29], the PlaceLab "Intensive Activity" Test Dataset [25] and the KTH actions dataset [34]. As mentioned before, the environment is controlled to facilitate the detection and tracking process. Except for the Keck dataset [29] and the KTH actions dataset [34], a multiple camera set-up is used to capture as many details as possible. To capture more images per time interval, the HumanEva [36] and the CMU-MMAC [4] database use a higher framerate to capture the video. Also the resolution of the datasets varies because of the choice of the cameras. For tracking, pose estimation and gesture recognition it is favorable to have a high resolution to capture the smallest details, a high framerate to have the smallest movements and multiple camera to see the object of interest from as many angles as possible. For gesture recognition it holds that the faster the gestures are, the more accurate the different camera views have to be synchronised. Since the benchmarks up to now perform rough gestures like waving, punching, etc. the requirements on the synchronisation are not that severe yet. However, if the gestures become faster like in a real fight, or very detailed like in sign language, the synchronisation should be done better. In the HumanEva [36] the synchronization is done by software and in the MuHAVi-MAS dataset [7] there is no explicit synchronization done at all. Table 2 gives an overview of the multicamera benchmarks for pose estimation and gesture recognition.

Another important difference in the datasets is how they provide additional information as ground truth information. Since pose recognition consists of detecting the positions of body parts, a logical choice is to use a motion capture system as a ground truth. The HumanEva [36] and the CMU-MMAC [4] benchmarks obtain motion capture (MoCap) data captured by a Vicon system [14]. This is an industry standard for optical marker-based motion capture. The system uses reflective markers and infrared cameras at 120 Hz to recover the 3D position of the markers and thereby estimate the 3D articulated pose of the body. The HumanEva dataset used six 1M cameras and the CMU-MMAC used twelve 4M cameras. The placements of the markers (typically 40-60) are positioned to measure the 3D position of the entire human body. This MoCap system has also been used for pedestrain tracking in the Stereo Pedestrian Detection Evaluation Dataset [26]. The main drawback of the state-of-the-art benchmarks for pose and gesture recognition including MoCap data is that they restrict theirselves to the pose and/or gestures of one person only. There are datasets that incorporate multiple persons in combination with MoCap data such as the CMU Graphics Lab Motion Capture Database [3], the data set used by Liu [30] and the Stereo Pedestrian Detection Evaluation Dataset [26].

Table 2: Properties of multicamera benchmarks for pose estimation and gesture recognition.

Name	Year	No. cameras	Frame rate	Resolution	No. subjects	No. frames	No. sequences	Ground truth
CMU-MoBo [24]	2001	6	30	640 × 480	25	200,000	100	-
IXMAS [38]	2006	5	23	390 × 291	11	-	39	reconstructed volumes
HumanEva [36]	2007	7[1]	60	640 × 480	4	80,000	56	Vicon MoCap data *12 cams, 195 mpp*[3]
CMU-MMAC [4]	2009	6	30-60	640 × 480 / 1024 × 768	43	-	-	Vicon MoCap data *16 cams, 40 mpp*
MuHAVi-MAS [7]	2009	8	25	720 × 576	14	-	119	Manual annotations
TUM Kitchen [37]	2009	4	25	384 × 288	4	-	20	Markerless MoCap data
MPI08 [33]	2010	8	40	1004 × 1004	-	24,000	54	3D laser scans
UMPM	2011	4[2]	50	644 × 484	30	400,000	36	Vicon MoCap data *14 cams, 37 mpp*

[1] The first HumanEva dataset is recorded with 4 color and 3 greyscale cameras. The second HumanEva dataset only uses the four color cameras.

[2] The 4 color cameras do not face each other directly to avoid similar sillouettes.

[3] *cams*: cameras and *mpp*: markers per person.

However, the first one provides only MoCap data and no video and is restricted to interaction between two persons, the second one provides MoCap data for one person only in a two person interaction scenario and the last one is aimed at pedestrians only. Some benchmarks include alternative, but similar information instead of MoCap data obtained by a Vicon system. For example, the MPI08 database [33] uses 3D laser scans of the human body. The TUM Kitchen Data Set [37] provides MoCap data by using their markerless motion capture results obtained by their own method [15].

In addition to the video streams most benchmarks include more data streams, like audio, internal measurement units and wearable devices. Audio can be used to determine when a door opens. In the SPEVI dataset [13] two microphones are used in combination to a single camera view to determine the location of people. The CMU-MMAC dataset [4], TUM Kitchen Data Set [37] and the PlaceLab ”Intensive Activity” Test Dataset [25] also use data coming from 3-axial accelerometers, gyroscopes, magnetic sensors and/or microphones to support and supply their video recordings.

Next to measurements of additional sensors, most benchmarks also provide manually added ground truth information. This can be the position on the ground floor at a certain time or the coordinates of a 3D bounding box around a person. In action/gesture detection, this ground truth can be a starting time of an event or a duration. For poses, it can be a manually annotated segmentation. The H3D dataset [16] provides manually annotated poses for their video recordings and a way to produce these annotations in a fast way. Another example is the Buffy Stickmen V2.1 dataset [22], where for each imaged person, line segments are provided indicating location, size and orientation of six body parts (head, torso, upper/lower right/left arms). Another manual anotation to facilitate body pose estimation is a manually foreground segmentation as is provided in the MuHAVi-MAS data [7] and the Keck Dataset [29].

To overcome the artificial laboratory settings, the Hollywood Human Actions Dataset [28] has been introduced for action recognition such as standing up or kissing. The recordings show a more natural environment. However, these are monocular recordings without any ground truth on a persons position. The Buffy Stickmen V2.1 dataset [22] is also an example of a natural environment, but with the associated ground-truth stickmen annotations provided for one person per frame. Since movies for television broadcasting are not multicamera yet, the benchmarks obtained in this way will remain monocular.

When people perform gestures, a significant class of gestures are performed by the hands. This is also shown by the presence of databases based on hand gestures like the Pointing and command gestures database [11] for a vocabulary to replace mouse buttons and movements by hand gestures, the Sign Language Pose Recognition database [17] for sign language recognition and the Cambridge Hand Gesture Dataset [27] for general hand movements. Although hand gestures form an important aspect, the focus on only hand gestures will not be sufficient for us.

3 Acquisition

This section describes how the UMPM benchmark has been acquired. First the hardware and software configuration (camera types, speed, resolution, etc.) used to capture the data is discussed in detail. Next, the requirements on the subjects and the scene are given. Finally, a description of the recorded scenarios is presented.

3.1 Hard- and software description

The UMPM benchmark captures both color video images and motion capture data simultaneously by using hardware synchronization. To have a compete insight in the acquisition of the data, the details of the color cameras, the motion capture system, and the hardware synchronization are given here.

Color cameras

To capture the video sequences, the room is equipped with 4 Basler PiA A640-210-gc color cameras with a resolution of 644 × 484 and a frame rate of maximal 210 fps[1]. The video output type is Gigabit Ethernet (GigE Vision compliant) and is triggered by an external trigger coming from the Ultranet HD (see hardware synchronization). The capturing speed

[1] See http://www.baslerweb.com/beitraege/unterbeitrag_en_48536.html for more specifications.

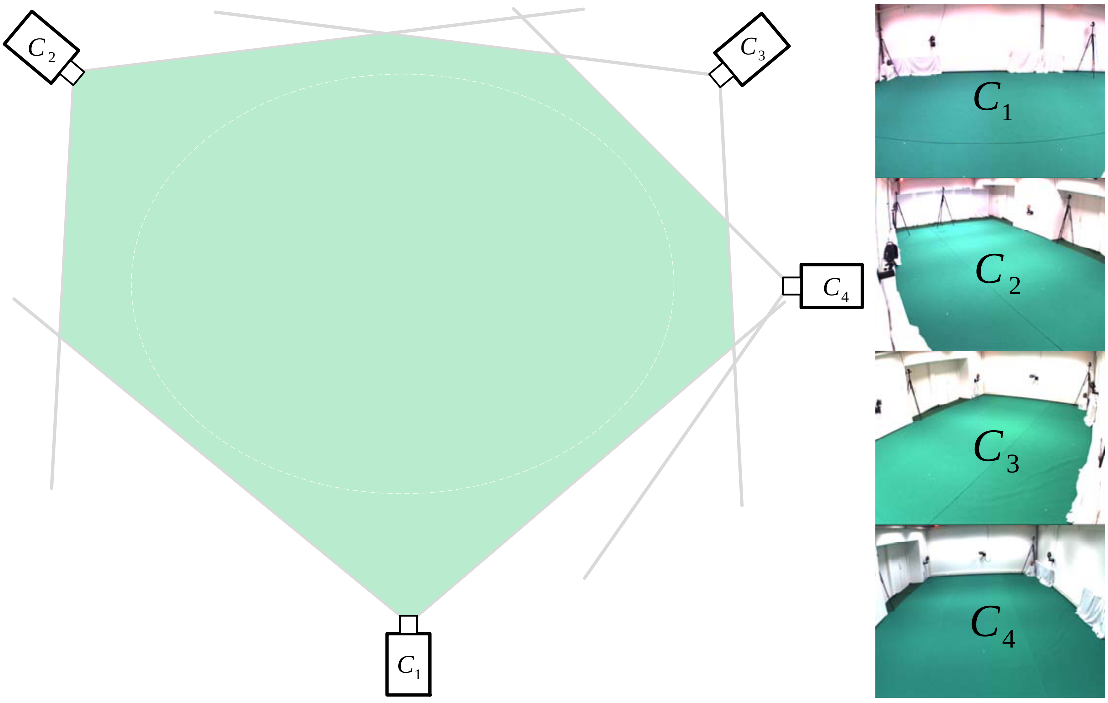


Figure 1: Top-view of the camera placement of the four monocular color cameras (*left*) and the camera views (*right*).

is chosen to be 50 fps. The cameras are placed such that 3 cameras form an equilateral triangle together with the fourth camera as shown in Figure 1. This choice ensures that (1) the cameras surround the acquisition area and the subjects, (2) there is sufficient overlap between the cameras field of views, and (3) the cameras do not face each other directly to avoid similar silhouettes. They are fixed about 2 meters above the ground on a tripod, focused at the center of the scene.

The cameras have a wide angle lens (3.5 mm) to capture wider views such that the subjects can perform closer to the camera, the size requirement of the scene is reduced and the camera placement setup is more flexible. Although a wide angle lens has much more radial distortion that a normal one, this can be well corrected for by software given the camera parameters after the calibration.

The settings of the camera are manually set by Pylon Viewer (version 2.2.1.2148), which is the accompanying software from Basler. This software is used to set the white-balance and the gain. By switching on the Ultranet (and thereby the cameras), the Pylon viewer is opened first. Each camera is selected to

- set **continuous grab**;
- do **gain auto once** in the Analog Controls;
- hold a blank sheet of white paper in front of the camera and do **balance white once**;
- push **stop grab**.

After this, the Pylon viewer is closed and the cameras are ready to record with the help of the Nexus software.

For methods that rely on background subtraction to detect or track a subject, we provide the background images, which are video sequences captured by the color cameras without any subject in the scene. For the scenarios where for example a chair or a table is present in the scene, we record the background once more.

To link the 2D images captured by the color cameras to each other, calibration must be performed. Since the cameras are fixed, the calibration process is only performed at the beginning of a recording sequence. To compute the calibration parameters for each camera, we move a large checkerboard in each camera view and these image sequences form the input of the calibration process. Remark that for setting up the benchmark, the calibration should be done as accurate as possible, meaning we have to move the checkerboard accross the whole view equally. To have a robust calibration tool, the checkerboard pattern was printed on a single A0 paper and tape this onto a light and strong board (see Figure 2 (*left*)). Section 4.2 explains how the calibration parameters are obtained from this image sequences.

The captured movies are stored in an uncompressed avi format and are directly available for download on the benchmark webpage http://www.projects.science.uu.nl/umpm/. Each movie will contain 30-45 seconds and varies in size from 2 to 4.5 GB. The files are compressed without loss with xz, which reduces their file size with a factor of about 4.

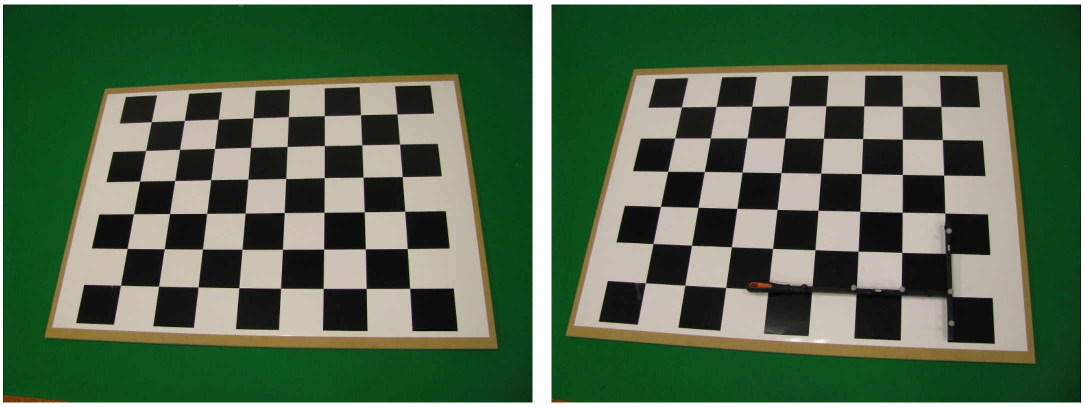


Figure 2: Picture of the calibration checkerboard (*left*) and the checkerboard with the wand placed on top of it to globally align all cameras (*right*).

Vicon system

To obtain the ground truth 3D data, the room is equipped with a Vicon MoCap system, which consists of 8 Vicon MX-40+ cameras (4 megapixel resolution, maximum speed of 160 fps, infrared), and 6 Vicon MX-F40 cameras (4 megapixel resolution, maximum speed of 370 fps, nearinfrared). See Figure 3 to see a picture of these cameras. The actual recording speed is set to 100 fps, which is twice the frame rate of the color cameras.

The 8 infrared cameras are positioned as high as possible on the wall, such that the color cameras can be focused on the scene without facing the infrared source directly. The 6 near-infrared are positioned lower, since they will not interfere with the video capturing and robust capturing of the markers demand cameras closer to the floor. To have a robust capturing of the MoCap data, the markers must be detectable by at least 2, preferably more cameras. Since the room is too small to focus all cameras simply at the center of the scene, the floor is divided into four quadrants and the cameras are divided into four groups to focus on each quadrant. Tests show that markers can be captured consistently over the scene and only occlusions of the subjects itself (self-occlusion), occlusions by other subjects and occlusions by static objects interfere with the capturing process of the invididual markers.

(a) Basler PiA A640-210-gc

(b) Vicon MX-F40

(c) Vicon MX-40+

Figure 3: Product images of the cameras used in the UMPM benchmark.

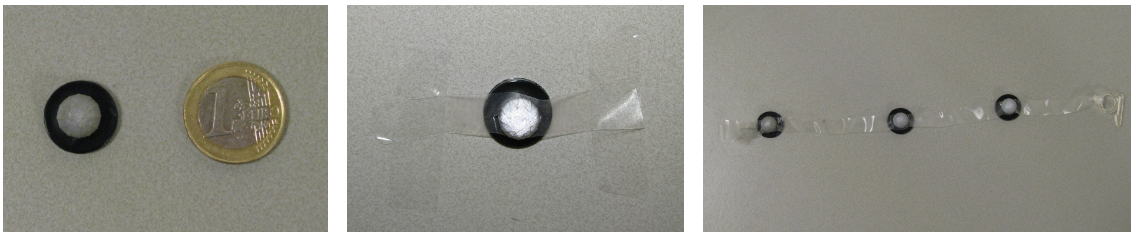


Figure 4: The 9.5mm reflective marker (*left*), the marker attached with tape (*middle*) and the marker attached on elastic brace (*right*).

The Vicon system identifies the 3D position of the reflective markers attached to the subjects. The markers used are 9.5 mm reflective markers which are hard threaded on a plastic base. They are attached to the subject by transparant elastic which is fixed to the subject by transparant tape if it concerns only one marker (like on the shoulder or the thigh) or it is put around a joint like an arm brace if more markers are used (see Figure 4). In this way, the appearance of the subject is kept as natural as possible.

To capture the MoCap data, the Vicon Nexus software (version 1.7.0.60305) is used. This software controls the recording and is used for processing the results. In the recording phase the video and motion capture cameras are selected and parameters can be adjusted such as the amount of infrared light emitted and the capture frequency for the Ultranet, the color cameras and the Vicon cameras.

In order to make the recording process as swift as possible, as much as possible is postponed to the postprocessing phase. The video data is recorded in a camera specific raw format. The Nexus software can transcode this later to a general format.

The information of the 3D marker positions are exported by the Vicon Nexus software in a C3D file format[2] This is a binary format, which stores the position of every label in every time frame.

Before the Vicon Nexus software can give 3D positions of the markers, the Vicon cameras

[2] See http://www.c3d.org/ for more details.

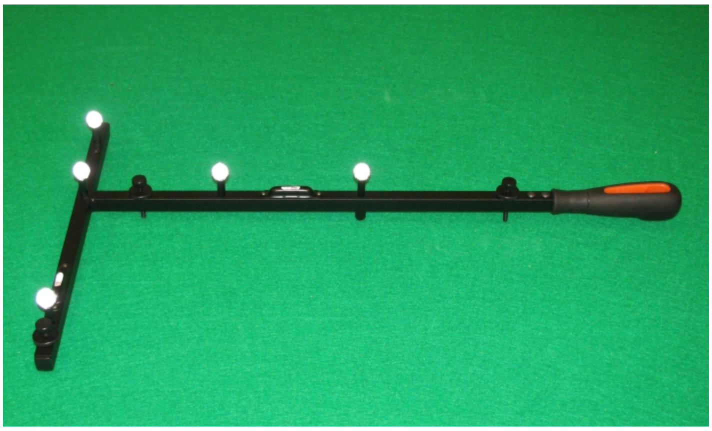


Figure 5: Image of the Vicon calibration wand.

need to be calibrated. The tool used for this is the calibration tool, which is a 5-marker wand, as shown in Figure 5. See http://www.cacs.louisiana.edu/labs/ecrg/vicon/index.html for more details. The calibration steps include

- making a mask for each camera, masking infrared sources that are not caused by markers (e.g., infrared lights from opposite cameras);
- waving the wand such that all cameras recognize it in sufficient places;
- placing the wand on top of the checkerboard in order to set the global origin at the right place (see Figure 2 (*right*)).

The last step is done to define a common origin.

Hardware synchronization

Both the color and Vicon cameras have been mutually synchronized by a hardware module of the Vicon MX Ultranet HD. See the diagram of the hardware connection in Figure 6. In our setup, we are limited to four synchronised video cameras, which is the maximum number the Ultranet HD supports.

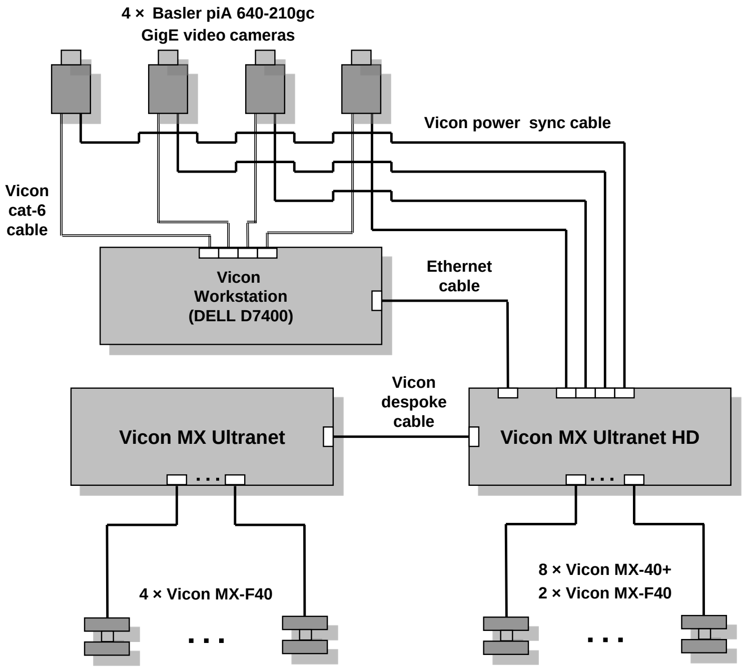


Figure 6: Diagram of the hardware setup and connections.

3.2 Requirements on subjects and scene

Besides the hardware and software, the quality of the benchmark is also determined by what is actually captured. Although we try to capture images that are representative for the real world, the subjects and the scene will be constrained to ensure that we capture what we have to capture.

Subjects

Before the requirements on the subject are explained, we state that there is no unlimited access in variety of persons. We are also depending on the availability of persons with a specific race, length, body size, etc. Therefore, it is of no use to make specific requirements as long as we have a reasonable set of persons representing the normal population. How the subjects dress, style their hair or use make-up is something we can prescribe. Some requirements are

- the hair style is natural and hair may not hide the face (eyes, mouth, etc.) and especially may not hide the reflective markers in any way;
- persons should wear natural clothes, but because of the marker attachments, the clothes will in general be tight (no loose clothes);
- the colors of the clothes and/or the pattern of colors should be distinguishable from the background to ensure proper background subtraction;
- the subjects may not wear glasses or other shiny objects (like watches, jewelry, etc.), because of interference with the Mocap data.

The markers are attached to subjects to capture the movements of the body and limbs. To face the challenge of inter/intra-person occlusions, we position the markers as illustrated in Figure 7. For each person we used 37 markers: 3 around the head; 2 around the neck; 4 around the waist; 1 on the shoulder; 3 around each elbow; 3 around each wrist; 1 on the outside of the hip; 3 around each knee; and 3 around each ankle. The positioning of the markers is customized to handle inter-person occlusions. Each joint of the human body (wrist, ankle, neck, etc.), except the shoulder and hip, is measured by more than one marker. For example, the wrist has three markers to indicate the center. If a marker is occluded, other markers for this joint might still be detected.

If only one subject is in the scene, this subject is fully equiped with markers. If multiple subjects are in the scene, no more than two subjects wear markers. In our opinion the added value of having more than two subjects wear markers is only marginal, but the challenge of process missing markers decreases exponentially as the number of fully equipped subjects increases. The developped algorithms to perform multi-person (articulated) motion tracking can still be validated by using the MoCap data for two persons. If the two subjects show nice results, the other persons will do nice as well. The diversity in videos in this benchmark ensure that the method is tested for variety in appearance, shape and motion.

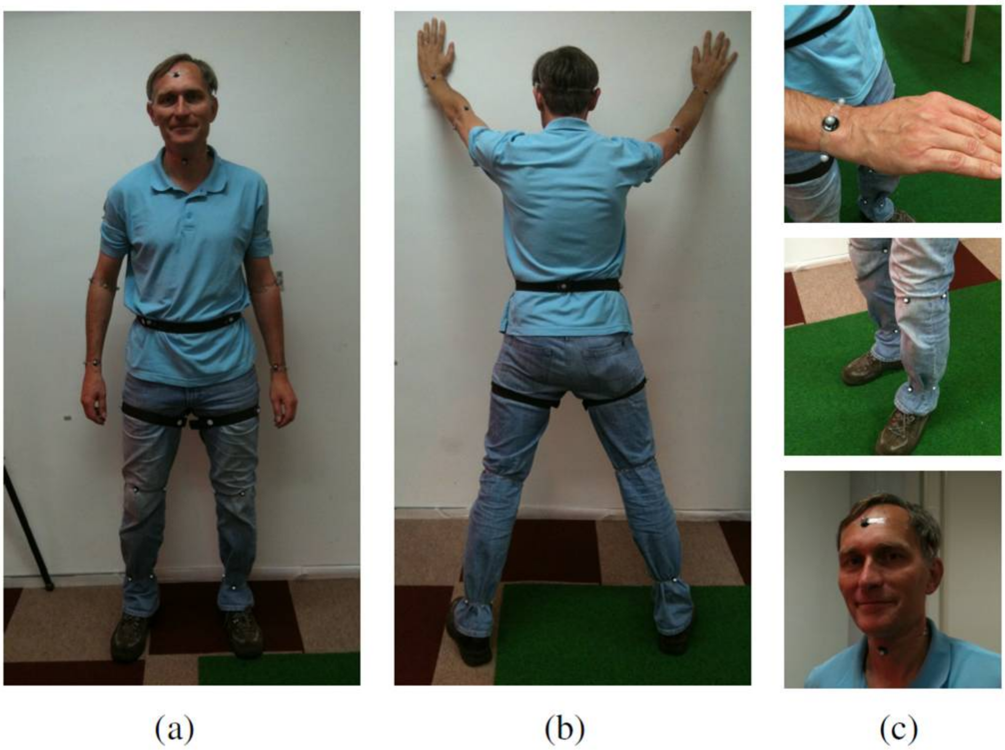


Figure 7: Placement of the reflective markers: frontal view (a), back view (b) and detailed views of the wrist, the knees and ankles, and the head and nek (c).

The participants are master students, PhD students and staff members of the Computing Science Department of Utrecht University. Participation in the collection process was voluntary and each subject was required to read, understand and sign a consent form for collection and distributation of data. A copy of the consent form is available by writing the authors. Subjects were informed that the data, including video images, would be made available to the research community and could appear in scientific publications.

Scene

To optimize the capturing of the reflective markers with the Vicon system, the benchmark is recorded in an indoor environment. Natural light is blocked and only artificial light in the lab is used to reduce the noise level of the Vicon system. The light will be influenced by the (near-)infrared red lights from the MoCap MX cameras.

The room has white walls and a green carpet covers the floor. Because of the size of the room and the positions of the Vicon cameras (fixed on the wall), the captured region is about 36 m^2 (6 m × 6 m) and the subjects are supposed to stay within this region. Some static objects like tables and chairs are possibly placed inside the scene depending on the scenario that is recorded. The choice of the carpet color is based on the fact that the subjects should be distinguishable from the background. Since the floor is green, and the wall is white, the subject should not wear these colors.

Additional efforts are made to ensure background subtraction can be performed. Cables are put into a shute and plaits cover the radiators to hide them from the camera. To hide the computers from the videos, blue movable walls are used to separate the system from the scene. See Figure 1 for an impression of the scene.

3.3 Scenario description

The body poses and gestures persons show can be classified according to

1. natural poses and gestures used in daily life;
2. synthetic poses and gestures, which are special human movements for some particular purpose, e.g. human-computer interaction, sports, body games etc.

To simulate such difficult cases for evaluating the algorithms, each of these two classifications of motions is subdivided into different scenarios, which are performed by human subjects equipped with MoCap markers and performed in a given area.

The scenarios of natural motions concern normal movements of humans in daily life such as walking, jogging, balancing, running, throwing etc. There is a big difference in the detection of gestures (or poses) that are commonly known and applied by people in ordinary situations and gestures that have to look natural to other people, but have to be identified by a camera to control for example a PowerPoint presentation. Most gestures we perform are carried out with the hands or the head. On http://en.wikipedia.org/wiki/List_of_gestures a list is presented with one-hand and two-hand gestures and gestures with other body parts. See also Table 6 and 7 in Appendix A.

Synthetic motions are special movements that people do not usually do in daily life, but pose for some particular purpose, e.g. human-computer interaction, sports, body games etc. See Table 8 of Appendix B for the synthetic poses used in our benchmark.

For natural motion we defined 5 different scenarios where the subjects (1) walk, jog and run in an arbitrary way among each other, (2) walk along a circle or triangle of a predetermined size, (3) walk around while one of them sits or hangs on a chair, (4) sit, lie, hang or stand on a table or walk around it, and (5) grab objects from a table. The objects are shown in Figure 8. These scenarios include individual actions, but the number of subjects moving around in the restricted area cause inter-person occlusions. We also include two scenarios with interaction between the subjects: (6) a conversation with natural gestures, and (7) the subjects throw or pass a ball to each other while walking around. The scenarios with synthetic motions include poses are performed when the subjects (8) stand still and (9) move around. These scenarios are recorded without any static occluders to focus only on inter-person occlusions. In Table 3 each scenario is described and labeled.

Each scenario is recorded with 1, 2, 3 or 4 persons. Before starting any actual action, all subjects perform a T-pose at their starting position (see Figure 9). In this pose all markers and body parts are maximally visible, which ensures a proper way to initialize the motion capturing. To check the synchronization of the video cameras, one person claps his hands

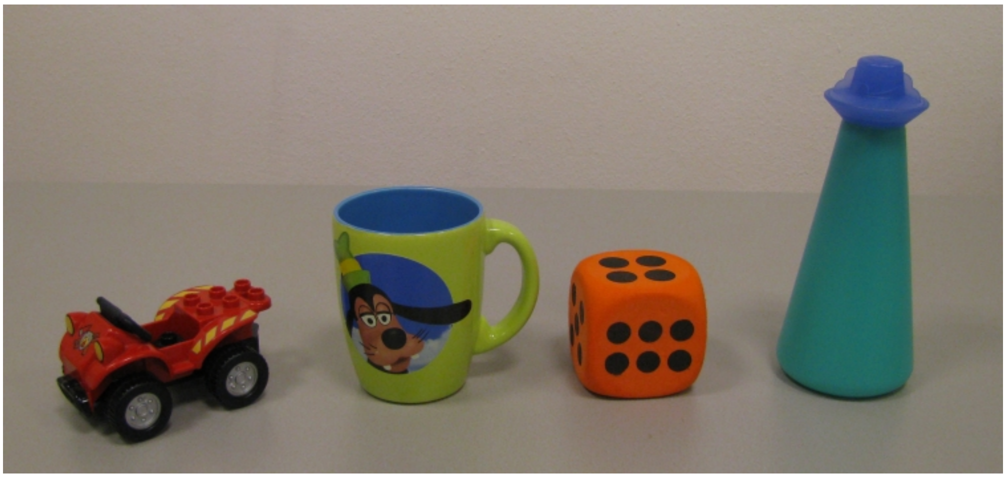


Figure 8: Objects used for the grabbing scenario.

before the scenario starts. At the end of each scenario, each subject returns to its starting position, one person claps again to obtain a second check of the synchronization, and each subject adopts the T-pose again. Each scenario will be recorded multiple times to provide variations within the same scenario, i.e. different subject combination, order of poses and movement patterns/trajectories.

An example of a file name is `p3_ball_2.c3d` which stands for a recording with 3 persons, scenario ball (or 2.2), take 2. The suffix tells that this is the C3D file.

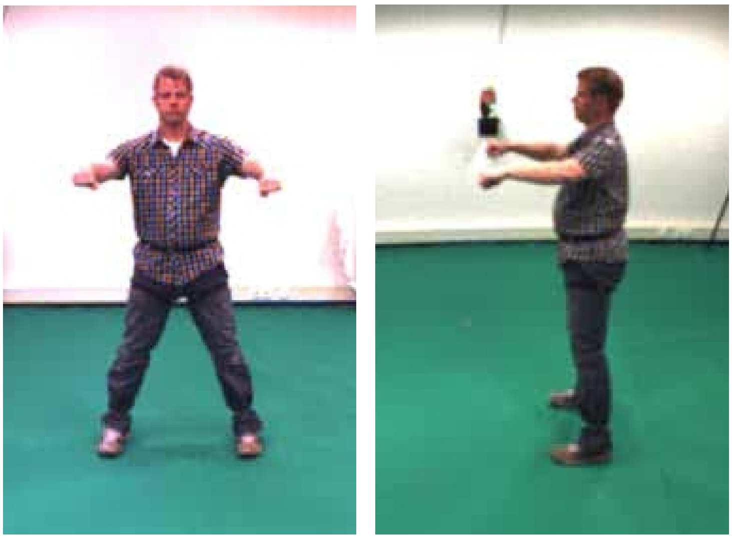
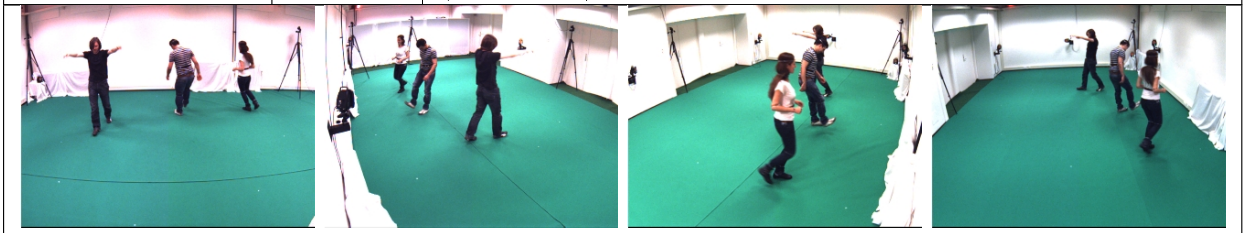
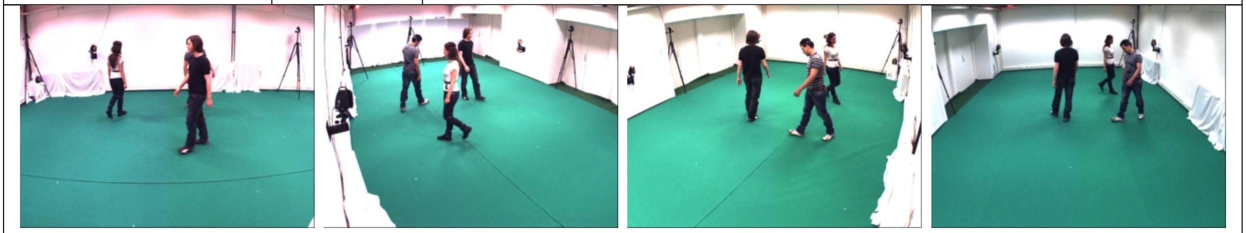
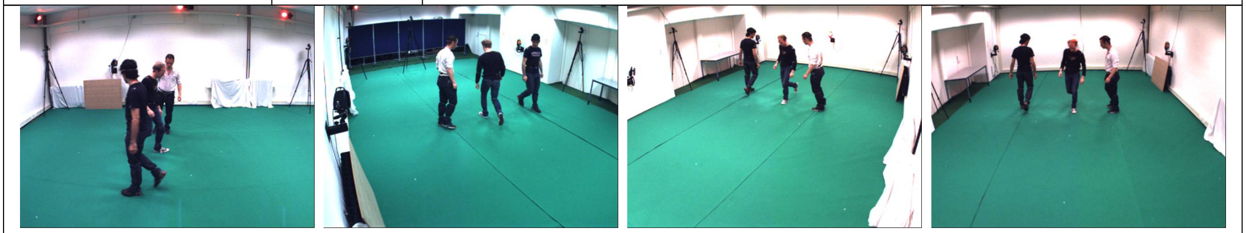
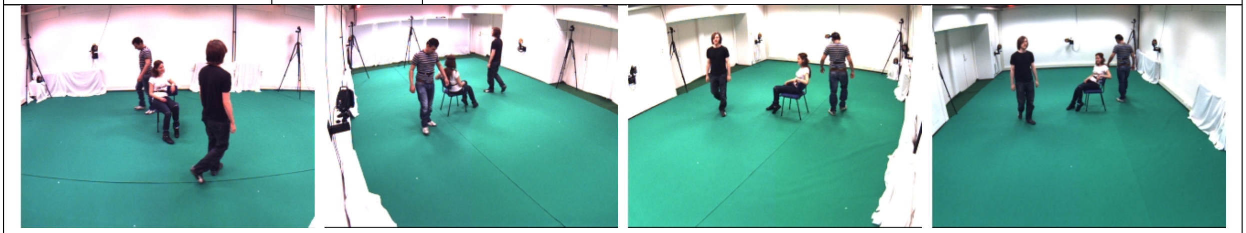


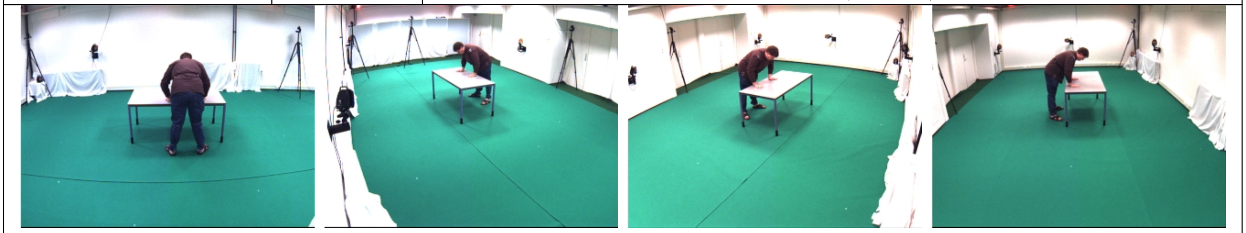
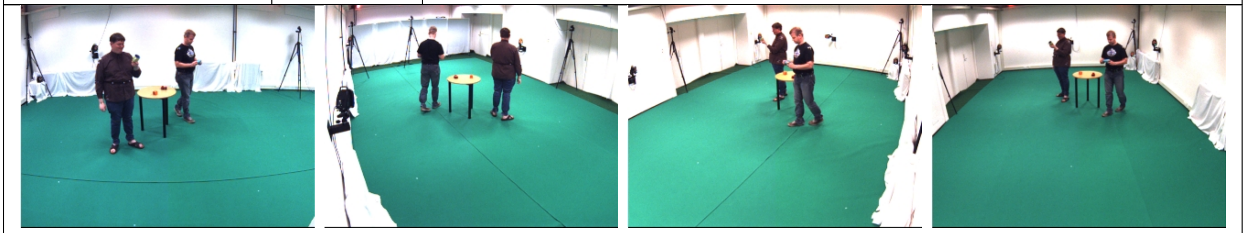
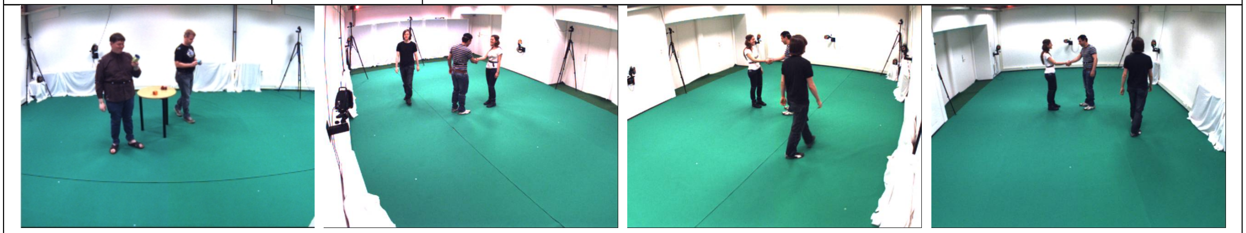
Figure 9: T-pose.

Table 3: Scenario play description.

Filename	**Subjects (n)**	**Description**
pn_free_k	1-4	The subjects start walking, running, jogging, balancing freely for a while arbitrarily independently from each other (45 seconds)
		
pn_circle_k	1-4	The subjects walk by following the path of the circle
		
pn_triangle_k	3	The subjects walk by following the path of a triangle within a circular area
		
pn_chair_k	1-4	The subjects start walking in a circle for a while, find the chair, stand still and sit on the chair, and stand up again, move and leave the chair to the next person.
		

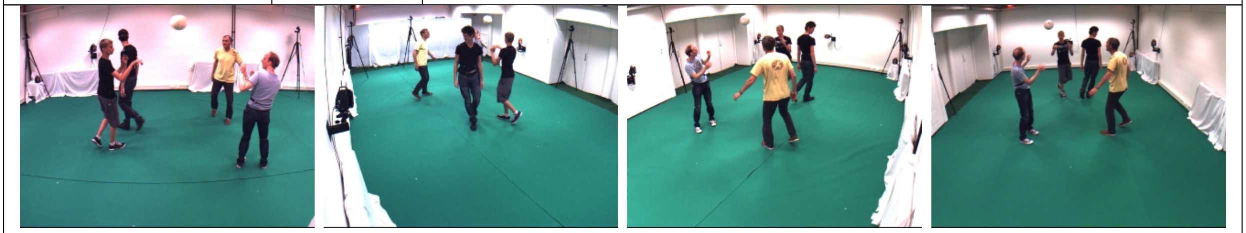
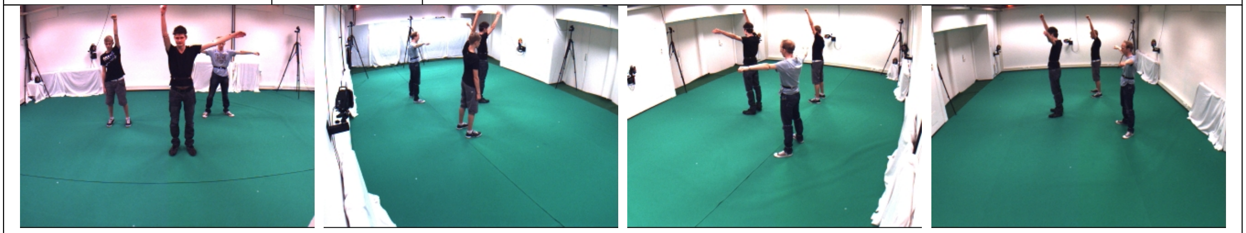
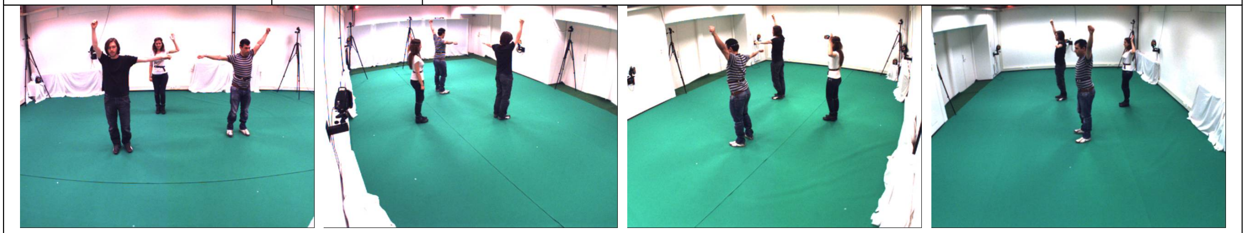
Continued on next page

Table 3 – continued from previous page

Filename	Subjects (n)	Description
pn_table_k	1-4	The subjects walk and lie down, lean against the table, bent forward. They can stand on the table and persons around it hanging on the table (or not)
		
pn_grab_k	1-4	On a table some static objects are present (which have been taken pictures of from multiple angles, light conditions, etc. to ensure the object can be perfectly identified). Each person grabs an object.
		
pn_meet_k	2-4	The subjects wave at each other as a sign of recognition, they give a handshake. During the fake conversation they nod or shake their head and make the shrug sign, shrug. The "conversation" will end with the beck sign where subject A gives something to subject B. As a response the person gives the OK sign. Both person give a handshake and leave.
		

Continued on next page

Table 3 – continued from previous page

Filename	Subjects (n)	Description
pn_ball_k	2-4	The subjects stand close to each other and start throwing a ball back and force between them, and at the same time constantly moving in a circle. 1 seconds interval between each throwing.
		
pn_staticsyn_k	1-4	The subjects start performing all the defined synthetic poses in arbitrary order, while holding their ground position. Normal posing speed, hold the pose for at least 1 second.
		
pn_orthosyn_k	1-4	The subjects start moving horizontally/vertically within the given area, and performing all the defined synthetic poses in arbitrary order. Normal posing speed, hold the pose for at least 1 second.
		

4 Benchmark data

The UMPM benchmark is located at http://www.projects.science.uu.nl/umpm/. This section describes the details of this benchmark, including the procedures of how to obtain this data from the captured data obtained in the acquisition phase to ensure the repeatability of this benchmark. The benchmark consists of the following data:

- color video sequences;
- calibration parameters of cameras;
- motion capture data C3D files.

The format of how this data is presented and how to use the accompanying software will be explained.

4.1 Color video sequences

The acquired video sequences from the four color cameras are available for download without any post-processing. The files are in raw avi format. To download all movies, the storage capacity should be 300 to 500 Gb.

4.2 Calibration

Camera calibration information is crucial information, since it relates the video cameras among each other and globally aligns the MoCap and video cameras. The calibration procedure of the Vicon cameras has already been explained in Section 3.1. The calibration parameters are computed by the Vicon Nexus software and used to provide the 3D position of the markers. Therefore, no calibration parameters are given in the benchmark for the Vicon cameras.

To calibrate the color video cameras from the checker-board recordings, the software tool of Zhang *et al.* [41] is used. This tool is publicly available in the OpenCV library[3] and returns the intrinsic/extrinsic and distortion parameters given the image sequences for each camera. Knowing the dimensions of the checkerboard in the 3D world and using a corner detector to find the intersections in the checkerboard pattern, the algorithm computes the desired parameters. Zhang's algorithm gives the following parameters for each camera, which we also provide in the UMPM benchmark:

- intrinsic parameters linking the pixel coordinates of an image point with the corresponding coordinates in the camera reference frame;
- extrinsic parameters defining the location and orientation of the camera reference frame with respect to a known world reference frame;
- radial distortion parameters.

[3] http://www.cs.iit.edu/ agam/cs512/lect-notes/opencv-intro/opencv-intro.html

To align the global origin to set the coordinate system that is shared by the Vicon and color cameras, the zero-coordinate of the ”wand” (for the Vicon cameras) and the checker-board (for the video cameras) are placed at the same spot on the ground floor (see Figure 5). This ensures that the 3D marker positions from the Vicon system are taken from the same 3D world as the 2D projections of the color cameras.

4.3 Motion capture data

The Nexus software combines the data from the Vicon cameras to obtain 3D marker positions. After manually assigning labels to the markers in one frame, the Nexus software makes a reconstruction of the trajectories and assigns labels to marker positions in the other frames. This reconstruction is not always correct due to frequent inter-person and intra-person (self) occlusions such that some of the reflective markers could not be seen by the Vicon system. This implies that (1) missing markers should be identified, (2) erroneous measurements should be removed, and (3) each marker should get a label of the body part and the person it belongs to. A missing marker is a marker that is not measured because this marker is not visible in a minimally required set of camera views due to e.g. occlusions. To have an insight in the quality of the UMPM benchmark concerning the measured markers, we give some statistics about these measurements for each recording. An example of these statistics is given in Table 4, but the full list will be available on the webpage http://www.projects.science.uu.nl/umpm/. This table shows the following statistics of each recording:

- total number of frames;
- total number of missing markers;
- average number of missing markers per frame;
- percentage of missing markers;
- maximum number of anonymous markers per frame;
- total number of anonymous markers;
- average number of anonymous markers per frame;

Note that these statistics are computed when one or two subjects wear the reflective markers. In each single person scenario, 37 markers are expected to be detected and in a multiple person scenario, this amount is doubled.

Manually checking half a million marker positions -a typical number in a recording- is a non-trivial task. We developed an approach to subsequently (1) check the continuity of trajectories, meaning that a sudden change in the speed of a label points to a place that needs special attention (this reduces the number of positions to inspect to about 100), (2) assign labels to marker positions that were not yet assigned a label (anonymous markers), where we iteratively check for discontinuities, relabel manually and label anonymous markers until no serious discontinuities are left, and finally, (3) interpolate trajectories and throw away anonymous markers. This procedure is repeated until no significant improvements are found.

Table 4: Statistics on measured MoCap data per scenario.

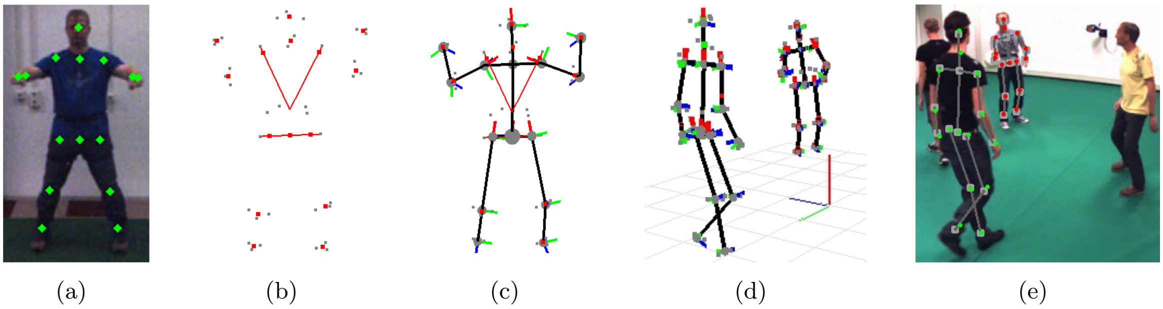
name	frames	total missing markers	average missing markers per frame	%mis	maximum anonymous markers	total anonymous markers	anony_fm
p2_ball_1.c3d	5591	7902	1.41	1.91	8	2803	0.50
p2_meet_1.c3d	4873	3837	0.79	1.06	14	11113	2.28
p4_ball_2.c3d	5942	7157	1.20	1.63	14	7830	1.32
p4_circle_1.c3d	6117	2037	0.33	0.45	9	3070	0.50
p4_free_1.c3d	6867	4552	0.66	0.90	13	8984	1.31
p4_meet_2.c3d	6422	8175	1.27	1.72	6	5234	0.82
p3_meet_2.c3d	5563	3249	0.58	0.79	13	6652	1.20
p2_grab_1.c3d	4820	21471	4.45	6.02	9	6783	1.41

The 3D marker positions, labels and the way these labels are found, are made available in the appropriate fields in a C3D file structure.

In the interpolation stage, we assume that all the marker's coordinates are known in the first time step and missing markers appear occasionally in the following time steps. We also assume that the markers attached on the same joint share similar motion while the subject moves, and the marker group's translations between frames are very similar to each other. Therefore the missing marker's current position can be estimated by the translation of the other visible markers on the same joint. For instance, the elbow has 3 markers to indicate a single elbow joint. If one or two elbow markers are missing, we first calculate the visible markers' translation between the previous and current time step, and then estimate the missing markers' current position by adding the previous coordinates of the missing markers with the translation.

Table 5: Marker names as used in C3D files

virtual marker	**real marker**	**description**
head	fhead rhead, lhead	head, front head
neck	fneck bneck	neck, front neck, back
pelvis	rasi, lasi rpsi, lpsi	anterior superior iliac spine posterior superior iliac spine
rshoulder, lshoulder	rshld, lshld	shoulder
relbow, lelbow	relbtop, lelbtop relbext, lelbext relblow, lelblow	elbow, top exterior of elbow elbow, low
rwrist, lwrist	rwrtop, lwrtop rwrlow, lwrlow rwrext, lwrext	wrist, top wrist, low wrist, exterior
rhip, lhip	rtopleg, ltopleg	outside of leg, near the hip
rknee, lknee	rkneefr, lkneefr rkneebk, lkneebk rkneeis, lkneeis	knee, front knee, outside at the back knee, inside
rankle, lankle	rankfr, lankfr rankbk, lankbk rankis, lankis	ankle, front ankle, outside at the back ankle, inside



(a) (b) (c) (d) (e)

Figure 10: Estimating the subjects' joints ground truth. (a) Example frame superimposed with virtual markers (green circles). (b) Estimating the virtual markers (green circles) by the 3D coordinates of the reflective markers (gray squares). (c) The kinematically constrained human skeleton model. (d) Virtual markers (colored squares) drive the skeletons (black bones and gray joints). (e) The virtual markers (colored circles) and the skeletons (gray lines and circles) superimposed on one example view.

Next to the 37 marker positions per subject, we also provide two sets with 15 virtual 3D positions to describe the bone joints (see Table 5 and Figure 10(a)). In the first set, the joint positions (except for the shoulders and tighs) are computed by averaging the coordinates of the corresponding markers, since these are placed around the joints. The corresponding mean is therefore "inside" the body part, which properly represents the bone joints. The left and right pelvis joints are approximated as points on the line between the two markers (gray points in Figure 10(b)) on the left and right upper legs, respectively. The points' positions are adjusted to match each subject's actual pelvis width, and the resulting virtual markers are also "inside" the torso (colored points in Figure 10(b)). The left and right shoulder joints are approximated as points on the line between the center of the waist and the two markers on the left and right shoulder tops, respectively. The points positions are also adjusted to match each subjects actual shoulder location (Figure 10(c)). Similarly, the center of the neck and the marker on the upper arm define the shoulder coordinates. Finally, the means of the pelvis joints and shoulder joints define the pelvis and chest centers, respectively.

The second set adds a kinematically constrained human skeleton model (23 degrees of freedom) to the 15 virtual joints of the first set to overcome the problem of moving markers caused by the deformation of muscles or moving clothes. It uses the Cyclic-Coordinate Descent (CCD)-based inverse kinematics approaches [39] (Figure 10(c)). The 3D coordinates of the virtual joints are therefore regarded as 'goal positions' to drive the skeleton (Figure 10(d)). The bone lengths of the skeleton are scaled according to the corresponding positions of the virtual joints in the first frame. The skeleton imposes strict poses and joint position constraints, which makes the estimation of the joint ground truth more robust against moving markers and measurement errors (10(c)).

On the benchmark webpage three formats of the ground truth 3D information will be available: (1) 37 marker positions per subject, (2) 15 joint positions obtained directly by averaging the marker positions, and (3) 15 corrected joint postions by enforcing kinematic constraints.

5 Future extensions

The UMPM benchmark has been introduced to evaluate algorithms for human motion capturing for multiple subjects in a similar way as HumanEva does for a single subject. Although HumanEva is broken down into three disjoint sub-sets: training, validation, and testing, we ”only” provide multiple sequences with different subjects of the same scenario. The user may choose to make such a distinction if desired.

Although we already provide the data to the research community, some additional efforts should be done to improve or demonstrate the quality of the data set. To help other researchers, some common algorithms could be implemented, or researchers of published papers could be asked to implement their algorithms on this data set. A webpage could be maintained where results of applying methods on this benchmark are registered. This will significantly save other researchers effort and increase the impact of this dataset.

The UMPM benchmark uses a heuristic method based on analysis of trajectory to recover the missed markers. More mathematical analysis provides a way to evaluate the performance of this method in terms of accuracy. To do that, some markers could synthetically be removed first and recovered later by using the proposed method. This is a possible way to compute the accuracy by comparing the estimated position with ground-truth.

The Motion Capture Lab of Utrecht University, where the UMPM benchmark has been recorded, is a permanent lab. This implies that the benchmark can be extended with other scenarios, specific subjects or objects in the scene, another illumination, different recording speed, etc. For more information please contact UMPM@science.uu.nl.

A List of commonly known gestures

In this appendix a set of examples gestures are given as you can find them in the scenario of `p`n`_meet_`k, where n is the number of persons and k is the take number.

A.1 Single hand signs

Many signs we give during interaction are done by the hands. Table 6 gives some examples.

Table 6: Single hand signs

Okay-sign	Beckoning sign	Handshake	Pointing
Thumbs up	Thumbs down	Wave	

Other gestures are done with the head alone or a combination of the head and the hand(s). In Table 7 some examples are given.

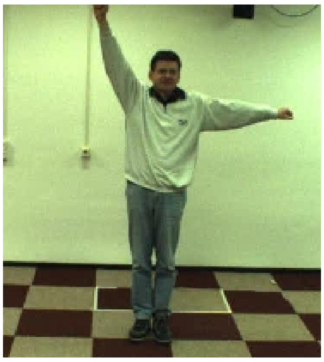
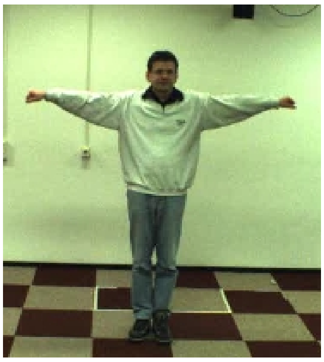
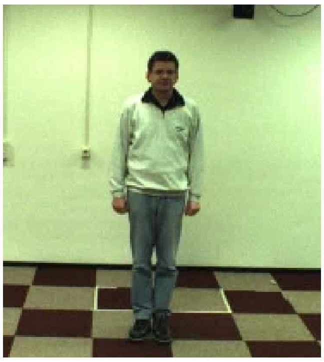
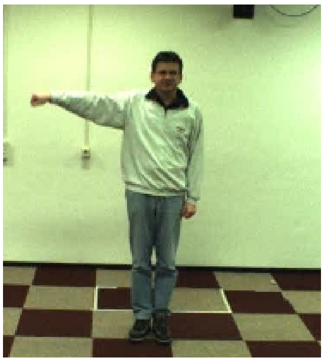
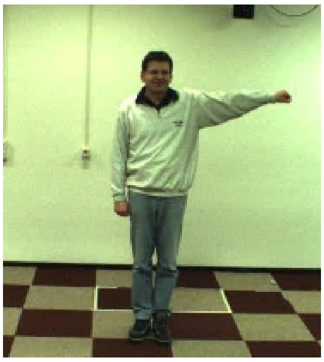
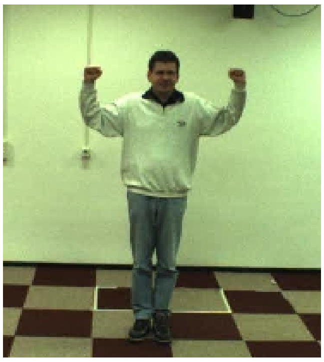
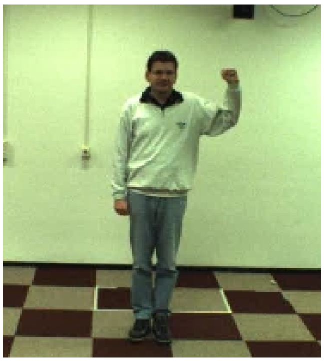
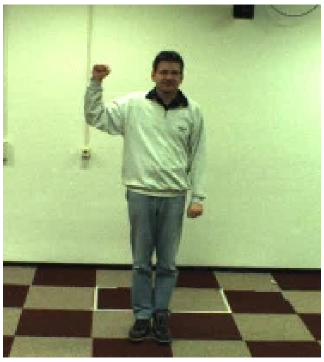
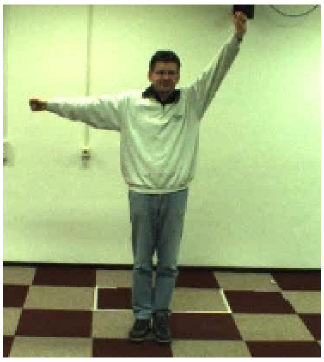
Table 7: Gestures with other body parts

Choking sign	Facepalm: sign to express frustration or embarrassment	Nod or shake head	Shrug, lifting both shoulders	Shush gesture

B List of synthetic poses

Table 8 gives an overview of the poses used in scenarios pn_staticsyn_k and pn_orthosyn_k. These poses are examples of non natural poses that might be used in human computer interaction.

Table 8: Synthetic poses

References

[1] BEHAVE interactions test case scenarios. http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/.

[2] CAVIAR test case scenarios. http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/.

[3] CMU motion graphics lab motion capture database. http://mocap.cs.cmu.edu/.

[4] CMU multi-modal activity database. http://kitchen.cs.cmu.edu/.

[5] Color feret database. face.nist.gov/colorferet/.

[6] Middlebury stereo datasets. vision.middlebury.edu/stereo/.

[7] MuHAVi-MAS. http://dipersec.king.ac.uk/MuHAVi-MAS/.

[8] PETS 2002. www-prima.inrialpes.fr/FGnet/data/09-Pets2002/pets2002-db.html.

[9] PETS 2006. www.cvg.rdg.ac.uk/PETS2006/index.html.

[10] PETS 2009. http://www.cvg.rdg.ac.uk/PETS2009/index.html.

[11] Pointing and command gestures dataset. http://www-prima.inrialpes.fr/FGnet/data/03-Pointing/index.html.

[12] REASON - PETS 2007. http://www.cvg.rdg.ac.uk/PETS2007/index.html.

[13] SPEVI datasets. http://www.eecs.qmul.ac.uk/ andrea/spevi.html.

[14] Vicon motion capture. www.vicon.com.

[15] J. Bandouch, F. Engstler, and M. Beetz. Accurate human motion capture using an ergonomics-based anthropometric human model. In *Proceedings of the Fifth International Conference on Articulated Motion and Deformable Objects (AMDO)*, 2008.

[16] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[17] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *BMVC*, 2008.

[18] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[19] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009.

[20] Markus Enzweiler and Dariu M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179 – 2195, October 2009.

[21] A. Ess, B. Leibe, K. Schindler, and L. van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008.

[22] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose search: retrieving people using their pose. In *CVPR*, 2009.

[23] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-camera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267 – 282, Februari 2008.

[24] R. Gross and J. Shi. The CMU Motion of Body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, July 2001.

[25] S. S. Intille, K. Larson, E. Munguia Tapia, J. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson. Using a live-in laboratory for ubiquitous computing research. In K. P. Fishkin, B. Schiele, P. Nixon, and A. Quigley, editors, *Proceedings of PERVASIVE 2006, vol. LNCS 3968*, pages 349 – 365. Berlin Heidelberg: Springer-Verlag, 2006.

[26] P. Kelly, N.E. O'Connor, and A.F. Smeaton. A framework for evaluating stereo-based pedestrian detection techniques. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1163–1167, August 2008.

[27] T-K. Kim, S-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *CVPR*, 2007.

[28] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[29] Zhe Lin, Zhuolin Jiang, and Larry S. Davis. Recognizing actions by shape-motion prototype trees. In *ICCV*, pages 444–451, 2009.

[30] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR*, 2011.

[31] M. Oren, C.P. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *CVPR*, pages 193–99, 1997.

[32] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson. A new pedestrian dataset for supervised learning. In *IEEE Intelligent Vehicles Symposium*, 2008.

[33] Gerard Pons-Moll, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *CVPR*, 2010.

[34] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[35] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4–27, 2010.

[36] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006.

[37] M. Tenorth, J. Bandouch, and M. Beetz. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE Inernational Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), 2009. In conjunction with ICCV 2009.*

[38] D. Weinland, R. Ronfarda, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 104(2-3):249–257, 2006.

[39] C. Welman. Inverse kinematics and geometric constraints for articulated figure manipulation. Master's thesis, Simon Fraser University, April 1993.

[40] Christian Wojek, Stefan Walk, and Bernt Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, 2009.

[41] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, November 2000.